Journal for Reproducibility in Neuroscience

Reproducibility of behavioral phenotypes in mouse models - a short history with critical and practical notes

Vootele Voikar

Neuroscience Center and Laboratory Animal Center, Helsinki Institute of Life Science

Correspondence:
Mustialankatu 1 G
Helsinki, Finland
vootele.voikar@helsinki.fi

Progress in pre-clinical research is built on reproducible findings, yet reproducibility has different dimensions and even meanings. Indeed, the terms reproducibility, repeatability, and replicability are often used interchangeably, although each has a distinct definition. Moreover, reproducibility can be discussed at the level of methods, analysis, results, or conclusions (1, 2). Despite these differences in definitions and dimensions, the main aim for an individual research group is the ability to develop new studies and hypotheses based on firm and reliable findings from previous experiments. In practice this wish is often difficult to accomplish. In this review, issues affecting reproducibility in the field of mouse behavioral phenotyping are discussed.

Crisis in reproducibility. Over the last ten years, the "reproducibility crisis" has often appeared in the headlines of scientific journals (3-6). Several factors have been identified as the major causes for irreproducibility – including p-hacking, cherry-picking, low statistical power, publication bias, and hypothesizing after results are known (7). However, these issues mostly occur after the animal experiments are done. There are many more items to consider during the planning and running of an experiment – good experimental design includes considerations regarding randomization, blinding, details of housing, husbandry and animal care, the definition of the experimental unit, inclusion and exclusion criteria, and the choice of animal subjects (the source, health status, strain, sex and age of the animal), among others. (8-10). Guidelines and recommendations (e.g. ARRIVE, PREPARE) are available for addressing these factors (11, 12). However, despite the fact that the ARRIVE guidelines have existed for ten years and are endorsed by more than 1000 journals so far, awareness of researchers and the quality of publications have not been sufficiently improved (13-15). In order to facilitate and enhance implementation of the ARRIVE guidelines, a revised version with exhaustive explanation and elaborative documentation was recently published (16, 17).

Paradigm shift. Mice and rats are the most widely used model animals in basic biomedicine. However, there has been a drastic change in the relative use of these two rodent species over time (Figure 1). Historically, the rat was the model of choice for behavioral studies but from the beginning of 1990s a sharp shift from rats to mice took place. Obviously, this was due to rapid technological development in genetic engineering and the ability to create genetically modified mice (i.e. transgenic or targeted mutants). Initially these mice were only available in the most advanced laboratories (18, 19), but within ten years the use of genetically modified mouse models was widespread. Importantly, as it became a routine tool for almost every team in biomedical research, there was likely a prevailing impression that behavioral assessment was the easiest part of the process in discovering the function(s) of each gene. Moreover, it was supposed that rat paradigms could be easily translated and applied to mice. Yet it quickly became clear that mice are not little rats, and extensive work with mice has serious challenges (20-22). Another caveat is that while rat behavior in the laboratory has been studied for decades with the clear goal of understanding the mechanisms of behavior, the mouse is in the majority of cases studied only in the context of genetic modification (phenotyping the effects of gene targeting). This means that we

DOI: https://doi.org/10.31885.jrn.1.2020.1375

may still be missing a lot of important basic information and knowledge about mouse behavior in laboratory conditions (note the trends in Figure 1).

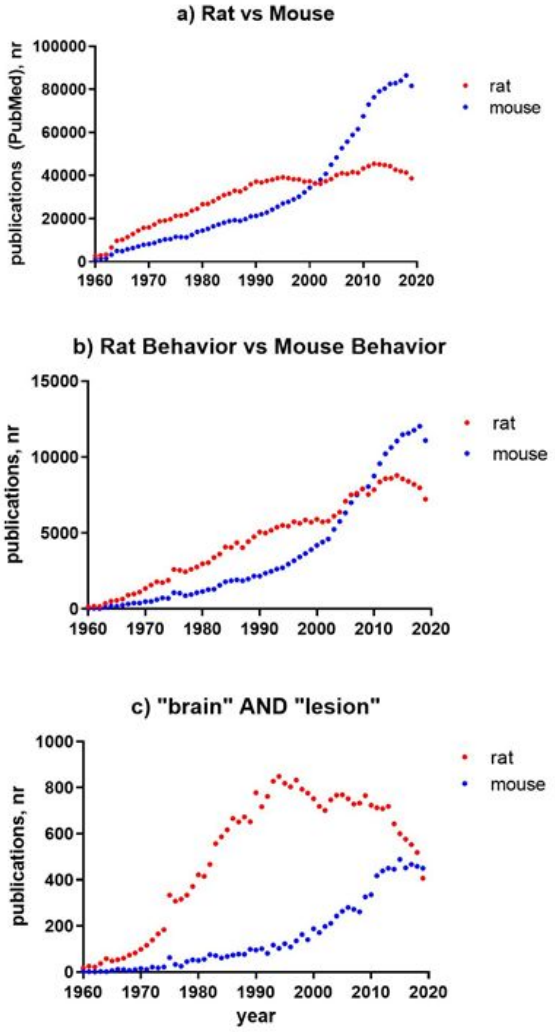

Figure 1. Simple PubMed search (accessed 5.6.2020) with keywords a) "rat" / "mouse" b) "rat and behavior" / "mouse and behavior" c) "rat and brain and lesion" / "mouse and brain and lesion".

A mouse is not a mouse. The first mutant mice were made using embryonic stem cells from the 129 mouse strain. However, it was known that these mice harbored several peculiarities complicating their use for neurobehavioral research, including poor breeding performance, hypo-activity, impaired learning, absent corpus callosum, and genetic contamination (23, 24). Therefore, the mice were crossed with another inbred strain, C57BL/6, which had been shown as a reasonable strain for various research topics, possessing intermediate phenotypes in many readouts that allowed identification of both gain- and loss-of-function (25). Subsequently, phenotypes of the mutant (knockout) and control (wild type) mice could be compared in the F2-hybrid generation. Yet the phenotypes of these F2 mice may be convoluted by the presence of unusual background genes, especially by flanking (passenger) genes from the 129 strain (26, 27). Thus, the recommendation was to continue backcrossing with parental strains until congenic and co-isogenic lines were established (28). According to this recommendation, researchers would have to have in hand two distinct genetic backgrounds with the possibility to make an F1-hybrid, which would have allowed for a powerful experimental design. However, in reality backcrossing has usually been done only to the C57BL/6J strain (considered a "gold standard" strain, after its genome was sequenced in 2001). In order to overcome the problems associated with a mixed genetic background and to facilitate the production of mutant mice, embryonic stem cell lines from the C57BL/6N strain were established (29). The International Mouse Phenotyping Consortium (IMPC) is currently creating mutant mice for large-scale phenotyping in a C57BL/6N background (30, 31). However, many researchers are unaware of the genetic and phenotypic differences between these two sub-strains (C57BL/6J and C57BL/6N), and the issue is further complicated by poor reporting of animal characteristics (full and correct strain name often missing) (32). Unfortunately, this is a major limitation for external validity and applicability of research with mutant mice. The use of inbred strains is well justified for reducing variability and increasing the precision of measurements (targeting genes in a known background). However, good design (and applicability) would require using more than one strain (33, 34) and there are numerous examples of phenotypic differences from the same mutation depending on the background strain (35).

The phenotype is a result of gene-environment interplay. If a pre-clinical research group has created a mutant mouse line, then eventually the phenotypic characterization of live animals will be conducted. The early literature of such studies

is full of controversies. Therefore, the behavioral neuroscience community has already been aware of the problems with (ir)reproducibility much earlier than the last decade, and in a way has been better prepared for the "crisis" (36-38). There has also been a major conflict between "molecular biologists" and "behaviorists" where the former could not understand or accept different results obtained by different laboratories. In order to tackle the discrepancies between laboratories, it was recommended to apply extensive standardization of procedures and environment. Such a solution was tested in a seminal study published in 1999 by Crabbe, Wahlsten and Dudek (39). They found that despite rigorous standardization of almost everything in three different laboratories, some of the results were idiosyncratic to a particular laboratory. Later, it was shown that among different factors contributing to the variability, the experimenter is the most prominent (40). However, there may be cases where standardization is required or desired – for instance, the IMPC has invested quite a bit in this type of effort, although the success has been variable (41). Exploring and tracking the causes for different outcomes between facilities may be a bumpy and painful process (42-44). All these findings exaggerated further the suspicions that behavioral studies are unreliable, an opinion especially expressed by researchers not working in the field of neurobehavioral research (43). However, an opposing theory was presented by Hanno Wurbel, suggesting that extreme standardization is a cause rather than cure for poor reproducibility (45, 46). Moreover, in addition to the principle of 3Rs (47), researchers working with animal models should adopt thinking in terms of 3V's – construct validity, internal validity and external validity (applicability, generalizability) (48). A comprehensive review of the current standing and future perspectives for embracing biological variation for enhancing reproducibility was recently published (49). Phenotyping efforts without considering the impact of environmental and developmental factors can be misleading.

Core facilities. Nowadays, making a knock-out mouse is a routine and standard procedure. The real challenge is in comprehensive phenotyping (50). In 2000, Jacqueline Crawley published a book, "What's Wrong With My Mouse" (second edition in 2007) (51), where she warned readers coming from molecular genetics that the aim was not to write a "how to" manual. Given that behavioral analysis is too complex to be treated as a "cookbook discipline", "descriptions of the methods are intentionally superficial", to give an overview of what is available. Inexperienced readers were advised to seek collaboration with experienced behavioral neuroscientists before setting up behavioral procedures. The real "pain and beauty" of mouse behavioral testing is comprehensively discussed in a book by Douglas Wahlsten (52). One solution for effectively tackling complexity in behavioral analysis has been the establishment of "core facilities" for behavioral assessment. By now, this is not surprising at all, because modern science is multidisciplinary, requiring special equipment (and more importantly, expertise) for adequately dealing with research questions. Thus, core facilities should help enhance the replicability and reliability of behavioral testing (53). The strengths and challenges of core facilities have been recently discussed (54, 55).

The need for establishing a core facility must be based on demands from the research community, the potential users of the facility. In Helsinki, the Transgenic Unit was formed at the Laboratory Animal Center in 1996 by the Institute of Biotechnology. Two years later, an initiative for behavioral analysis of mutant mice was launched and I was recruited for this purpose. The laboratory in Helsinki was developed from the beginning with the idea of being open and offering as broad support as possible to everyone interested in behavioral phenotyping – testing of basic sensory and motor functions followed by more complex tasks for coping with stress, learning and memory, and testing approach / avoidance behavior (e.g. exploration-curiosity, fear-anxiety).

However, as it often happens, I started with an instance of failure, which taught me a lot. The first transgenic line to be tested did not seem to

learn spatial navigation task in the water maze (at that time considered a gold standard test for learning). I then found out that the mice were in FVB/N background (standard strain for transgenic mice at that time) which suffer from a mutation causing retinal degeneration and blindness (56). Curiously enough, there were papers published showing spatial learning in this strain (57). This was a lesson that warned me that meaningful work with mutant mice requires parallel studies with inbred strains – know thy mouse (58-60) and methods (54)!

The end of 1990s was a very active period in the field of behavioral phenotyping – a new opening and shifting of paradigms. We learned a lot about differences between mouse strains (25) and about strategies to set up test batteries (as opposed to the tradition of conducting only one test per animal) (61, 62). Excellent international training courses and workshops were organized and I personally had the opportunity to attend courses arranged by Cold Spring Harbor Laboratory, EMBO / FENS, IBRO, IBANGS, and EUMORPHIA program. Inspiring interaction with fellow students and respected faculty (e.g. Jacqueline Crawley, Richard Paylor, Howard Eichenbaum, Seth Grant, Richard Morris, Hans-Peter Lipp, David Wolfer, Wim Crusio and many others) created a solid network and offered many good ideas for proceeding. My personal impression is that during the last 15 years there has been a decline in such high quality interactive courses. We have tried to fill this gap by organizing Baltic summer schools on the topic of rodent behavioral analysis (63). Indeed, learning, teaching and challenging existing paradigms are best achieved in communication and interaction between established and starting researchers – for impressions from recent FENS courses, see (64).

Quality monitoring! The main purpose of the core facility is to serve the research community (55). The organization of the facility and collaboration with users can have different forms, from full to minimum service. The responsibility of the facility is to maintain and take good care of the equipment and space, including monitoring of performance, necessary calibrations, and timely repairs and replacements. An essential part of the responsibility is training users and supporting them in all steps of the project (planning, conducting, analyzing, and reporting). This is one part of quality. Another part, as mentioned above, is to be up-to-date with developments in theory and methodology in behavioral neuroscience and laboratory animal science. In addition, internal validity and consistency should be verified with some standards and calibrations for animal behavior.

Users frequently ask what normal mouse behavior is, while thinking only in terms of their particular disease model. Moreover, it is often thought that the control, "wild-type" for gene-targeted mice, represents "normal", immediately implying that gene targeting will result in "abnormal" animals. As described previously, it is difficult or impossible to answer the question of what "normal" behavior is given the many inbred strains available, along with the impact of environmental conditions. Each inbred strain has some peculiarities due to inbreeding – retinal degeneration, deafness, anatomical differences, susceptibility or resistance for certain conditions that can develop (e.g. diabetes) (65). We all may have heard the saying "your genetics is only as good as your phenotype" and "this mutant does not have a phenotype" (66). First, these ideas may cause bias towards the hypothesis and second, having no phenotype is impossible – the only conclusion in this case would be that the given mice in the given situation did not display a phenotypic difference from the other study groups. Therefore, each model needs to be placed in the broader context of mouse behavior and with observation of the environment. Another question that may be asked is which is the best test for memory (or anxiety, or any other domain). Researchers with different backgrounds may not be aware of different memory systems or different types of anxiety. This is further complicated by the fact that there are hundreds of tests available (67). Navigating this landscape is one of the tasks of experts working in core facilities. Of course, the facility also learns from its users – a good

facility is open and flexible to adapting and developing new methods.

Quality monitoring can be done by regular testing of inbred strains with known phenotypes. Although we cannot speak of animals as tools, we hope that the phenotypes of inbred strains are stable over time (68). Therefore, the testing of such animals could reveal if the conditions in the laboratory are stable. The C57BL/6 and DBA/2 strains are the oldest strains available and much information has been collected about physiology, anatomy and behavior in these mice. In our studies, we have found consistent differences between these two strains in several conventional tests (open field, light-dark box, and forced swim test) throughout the years when our laboratory was located in three different buildings (60, 69-71).

Another important issue that is frequently discussed is the use of male and/or female mice (72). Indeed, it might be still difficult to convince researchers that including female mice in your studies does not ruin it – despite the evidence that female mice are no more variable than males (73, 74). Including both sexes is mandatory for sound design and enhanced external validity (75). Despite many years of recommendations to consider sex as a biological variable, the change is taking place very slowly (76, 77).

Finally yet importantly, the human factor (experimenter) in animal experiments cannot be neglected (40, 78). Therefore, handling techniques need to be trained and refined (79) in addition to improving the understanding of the behavior that is measured and recorded, even if the process is automated (80, 81). Testing animal behavior can be challenging – "Despite our best efforts, the mice will continue to win some innings" (82). For enhanced validity, I believe we need to put even more emphasis on recording and mining the behavior of animals in their home-cage environment (83). Nevertheless, it is not yet possible to fully replace conventional testing, yet at the same time the role of experimenter in the process of data collection needs to be critically evaluated (84).

Summary. In this short review, I tried to highlight the factors that I consider important or essential in running a meaningful (mouse) behavioral phenotyping program. I would like to conclude with the words of Michael Festing and Ulrich Dirnagl:

> 'We are not born knowing how to design and analyze scientific experiments (85).'
>
> 'We should be [moving to the world] where biological thinking rules, and sound data production is emphasized through careful planning, design, execution, and reporting of our studies. A world where methods and results are transparently described so that effects and inferences can be independently confirmed (86).'

Thus, I encourage everyone to be open to new ideas while being skeptical of the phrase "we've always done it like this".

Acknowledgements. Vootele Voikar is supported by Jane and Aatos Erkko Foundation. Professor David P. Wolfer (University of Zurich) is acknowledged for discussion and feedback on the draft.

References

1. Kenett RS, Shmueli G. Clarifying the terminology that describes scientific reproducibility. Nat Methods. 2015;12(8):699.
2. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med. 2016;8(341):341ps12.
3. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012;483(7391):531-3.
4. Begley CG. Six red flags for suspect work. Nature. 2013;497(7450):433-4.
5. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circulation research. 2015;116(1):116-26.

6. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533(7604):452-4.
7. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017;1:0021.
8. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS ONE. 2009;4(11):e7824.
9. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? PLoS Med. 2010;7(3):e1000245.
10. Smith AJ, Lilley E. The Role of the Three Rs in Improving the Planning and Reproducibility of Animal Experiments. Animals. 2019;9(11).
11. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 2010;8(6):e1000412.
12. Smith AJ, Clutton RE, Lilley E, Hansen KEA, Brattelid T. PREPARE: guidelines for planning animal research and testing. Lab Anim. 2018;52(2):135-41.
13. Menke J, Roelandse M, Ozyurt B, Martone M, Bandrowski A. Rigor and Transparency Index, a new metric of quality for assessing biological and medical science methods. bioRxiv. 2020:2020.01.15.908111.
14. Reichlin TS, Vogt L, Wurbel H. The Researchers' View of Scientific Rigor-Survey on the Conduct and Reporting of In Vivo Research. PLoS One. 2016;11(12):e0165999.
15. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Altman DG, Avey MT, et al. Revision of the ARRIVE guidelines: rationale and scope. BMJ Open Science. 2018;2(1).
16. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLOS Biology. 2020;18(7):e3000410.
17. Percie du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLOS Biology. 2020;18(7):e3000411.
18. Silva AJ, Paylor R, Wehner JM, Tonegawa S. Impaired spatial learning in alpha-calcium-calmodulin kinase II mutant mice. Science. 1992;257(5067):206-11.
19. Grant SG, O'Dell TJ, Karl KA, Stein PL, Soriano P, Kandel ER. Impaired long-term potentiation, spatial learning, and hippocampal development in fyn mutant mice. Science. 1992;258(5090):1903-10.
20. Gerlai R, Clayton NS. Analysing hippocampal function in transgenic mice: an ethological perspective. Trends Neurosci. 1999;22(2):47-51.
21. Whishaw IQ, Tomie JA. Of mice and mazes: similarities between mice and rats on dry land but not water mazes. Physiol Behav. 1996;60(5):1191-7.
22. Whishaw IQ, Metz GA, Kolb B, Pellis SM. Accelerated nervous system development contributes to behavioral efficiency in the laboratory mouse: a behavioral review and theoretical proposal. Dev Psychobiol. 2001;39(3):151-70.
23. Livy DJ, Wahlsten D. Tests of genetic allelism between four inbred mouse strains with absent corpus callosum. J Hered. 1991;82(6):459-64.
24. Simpson EM, Linder CC, Sargent EE, Davisson MT, Mobraaten LE, Sharp JJ. Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. Nat Genet. 1997;16(1):19-27.
25. Crawley JN, Belknap JK, Collins A, Crabbe JC, Frankel W, Henderson N, et al. Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. Psychopharmacology (Berl). 1997;132(2):107-24.
26. Gerlai R. Gene-targeting studies of mammalian behavior: is it the mutation or the background genotype? Trends Neurosci. 1996;19(5):177-81.
27. Crusio WE. Flanking gene and genetic background problems in genetically manipulated mice. Biol Psychiatry. 2004;56(6):381-5.
28. Silva AJ, Simpson EM, Takahashi JS, Lipp HP, Nakanishi S, Wehner JM, et al. Mutant mice and neuroscience: recommendations concerning genetic background. Banbury Conference on genetic background in mice. Neuron. 1997;19(4):755-9.
29. Pettitt SJ, Liang Q, Rairdan XY, Moran JL, Prosser HM, Beier DR, et al. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. Nat Methods. 2009;6(7):493-5.
30. Simon MM, Greenaway S, White JK, Fuchs H, Gailus-Durner V, Wells S, et al. A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. Genome Biol. 2013;14(7):R82.
31. Fontaine DA, Davis DB. Attention to Background Strain Is Essential for Metabolic Research: C57BL/6 and the International Knockout Mouse Consortium. Diabetes. 2016;65(1):25-33.
32. Ahlgren J, Voikar V. Experiments done in Black-6 mice: what does it mean? Lab animal. 2019;48(6):171-80.
33. Festing MF. Inbred strains should replace outbred stocks in toxicology, safety testing, and drug development. Toxicologic pathology. 2010;38(5):681-90.
34. Festing MF. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. ILAR J. 2014;55(3):399-404.
35. Sittig LJ, Carbonetto P, Engel KA, Krauss KS, Barrios-Camacho CM, Palmer AA. Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. Neuron. 2016;91(6):1253-9.
36. Bespalov A, Steckler T. Lacking quality in research: Is behavioral neuroscience affected more than other areas of biomedical science? J Neurosci Methods. 2018;300:4-9.
37. Kafkafi N, Agassi J, Chesler EJ, Crabbe JC, Crusio WE, Eilam D, et al. Reproducibility and replicability of rodent phenotyping in preclinical studies. Neurosci Biobehav Rev. 2018;87:218-32.
38. Crabbe JC. Reproducibility of Experiments with Laboratory Animals: What Should We Do Now? Alcohol Clin Exp Res. 2016;40(11):2305-8.
39. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science. 1999;284(5420):1670-2.
40. Chesler EJ, Wilson SG, Lariviere WR, Rodriguez-Zas SL, Mogil JS. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. Neurosci Biobehav Rev. 2002;26(8):907-23.

41. Mandillo S, Tucci V, Holter SM, Meziane H, Banchaabouchi MA, Kallnik M, et al. Reliability, robustness, and reproducibility in mouse behavioral phenotyping: a cross-laboratory study. Physiol Genomics. 2008;34(3):243-55.
42. Wahlsten D. Standardizing tests of mouse behavior: Reasons, recommendations, and reality. Physiol Behav. 2001;73(5):695-704.
43. Wahlsten D, Metten P, Phillips TJ, Boehm SL, Burkhart-Kasch S, Dorow J, et al. Different data from different labs: lessons from studies of gene-environment interaction. J Neurobiol. 2003;54(1):283-311.
44. Andrews AM, Cheng X, Altieri SC, Yang H. Bad Behavior: Improving Reproducibility in Behavior Testing. ACS chemical neuroscience. 2018;9(8):1904-6.
45. Richter SH, Garner JP, Wurbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods. 2009;6(4):257-61.
46. Wurbel H. Behaviour and the standardization fallacy. Nat Genet. 2000;26(3):263.
47. Tannenbaum J, Bennett BT. Russell and Burch's 3Rs then and now: the need for clarity in definition and purpose. Journal of the American Association for Laboratory Animal Science : JAALAS. 2015;54(2):120-32.
48. Wurbel H. More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. Lab animal. 2017;46(4):164-6.
49. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in light of biological variation. Nature Reviews Neuroscience. 2020.
50. Brown SD, Hancock JM, Gates H. Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. PLoS Genet. 2006;2(8):e118.
51. Crawley JN. What's Wrong With My Mouse? Behavioral Phenotyping of Transgenic and Knockout Mice. New York: Wiley-Liss; 2000.
52. Wahlsten D. Mouse Behavioral Testing. How to use mice in behavioral neuroscience: Academice Press; 2011.
53. Wahlsten D, Crabbe JC. Replicability and reliability of behavioral tests. In: Crusio WE, Sluyter F, Gerlai R, Pietropaolo S, editors. Behavioral Genetics of the Mouse Volume I Genetics of Behavioral Phenotypes: Cambridge University Press; 2013.
54. Gulinello M, Mitchell HA, Chang Q, Timothy O'Brien W, Zhou Z, Abel T, et al. Rigor and reproducibility in rodent behavioral research. Neurobiol Learn Mem. 2019;165:106780.
55. Bikovski L, Robinson L, Konradsson-Geuken A, Kullander K, Viereckel T, Winberg S, et al. Lessons, insights and newly developed tools emerging from behavioral phenotyping core facilities. J Neurosci Methods. 2020;334:108597.
56. Taketo M, Schroeder AC, Mobraaten LE, Gunning KB, Hanten G, Fox RR, et al. FVB/N: an inbred mouse strain preferable for transgenic analyses. Proc Natl Acad Sci U S A. 1991;88(6):2065-9.
57. Björklund M, Sirviö J, Puoliväli J, Sallinen J, Jäkälä P, Scheinin M, et al. Alpha2C-adrenoceptor-overexpressing mice are impaired in executing nonspatial and spatial escape strategies. Mol Pharmacol. 1998;54(3):569-76.
58. Voikar V, Koks S, Vasar E, Rauvala H. Strain and gender differences in the behavior of mouse lines commonly used in transgenic studies. Physiol Behav. 2001;72(1-2):271-81.
59. Voikar V, Vasar E, Rauvala H. Behavioral alterations induced by repeated testing in C57BL/6J and 129S2/Sv mice: implications for phenotyping screens. Genes Brain Behav. 2004;3(1):27-38.
60. Voikar V, Polus A, Vasar E, Rauvala H. Long-term individual housing in C57BL/6J and DBA/2 mice: assessment of behavioral consequences. Genes Brain Behav. 2005;4(4):240-52.
61. Crawley JN, Paylor R. A proposed test battery and constellations of specific behavioral paradigms to investigate the behavioral phenotypes of transgenic and knockout mice. Horm Behav. 1997;31(3):197-211.
62. Crawley JN. Behavioral phenotyping of transgenic and knockout mice: experimental design and evaluation of general health, sensory functions, motor abilities, and specific behavioral tests. Brain Res. 1999;835(1):18-26.
63. Dere E, Jolkkonen J, Voikar V, Tanila H. Editorial to the Special Issue: Animal Model of the Year 2036: Novel Perspectives in Behavioral Neuroscience. Behav Brain Res. 2018;352:1.
64. FENS/NENS schools [Available from: https://www.fens.org/Training/Training-Grants-and-Stipends/NENS-Grants/Slots-in-NENS-courses-or-programmes/What-they-say-about-the-Slots-in-NENS-courses/.
65. Stevens JC, Banks GT, Festing MF, Fisher EM. Quiet mutations in inbred strains of mice. Trends Mol Med. 2007;13(12):512-9.
66. Crusio WE. 'My mouse has no phenotype'. Genes Brain Behav. 2002;1(2):71.
67. Hanell A, Marklund N. Structured evaluation of rodent behavioral tests used in drug discovery research. Front Behav Neurosci. 2014;8:252.
68. Wahlsten D, Bachmanov A, Finn DA, Crabbe JC. Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades. Proc Natl Acad Sci U S A. 2006;103:16364-9.
69. Kulesskaya N, Karpova NN, Ma L, Tian L, Voikar V. Mixed housing with DBA/2 mice induces stress in C57BL/6 mice: implications for interventions based on social enrichment. Front Behav Neurosci. 2014;8:257.
70. Kulesskaya N, Voikar V. Assessment of mouse anxiety-like behaviour in the light-dark box and open-field arena: Role of equipment and procedure. Physiol Behav. 2014;133:30-8.
71. Ahlgren J, Voikar V. Housing mice in the individually ventilated or open cages-Does it matter for behavioral phenotype? Genes Brain Behav. 2019;18(7):e12564.
72. Beery AK, Zucker I. Sex bias in neuroscience and biomedical research. Neurosci Biobehav Rev. 2011;35(3):565-72.
73. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. Neurosci Biobehav Rev. 2014;40:1-5.
74. Fritz AK, Amrein I, Wolfer DP. Similar reliability and equivalent performance of female and male mice in the open field and water-maze place navigation task. Am J Med Genet C Semin Med Genet. 2017;175(3):380-91.
75. Clayton JA. Applying the new SABV (sex as a biological variable) policy to research and clinical care. Physiol Behav. 2018;187:2-5.
76. Karp NA, Reavey N. Sex bias in preclinical research and an exploration of how to change the status quo. Br J Pharmacol. 2018.

77. Mogil JS. Qualitative sex differences in pain processing: emerging evidence of a biased literature. Nature Reviews Neuroscience. 2020.
78. Bohlen M, Hayes ER, Bohlen B, Bailoo J, Crabbe JC, Wahlsten D. Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol. Behav Brain Res. 2014;272:46-54.
79. Hurst JL, West RS. Taming anxiety in laboratory mice. Nat Methods. 2010;7:825-6.
80. Stanford SC. The Open Field Test: reinventing the wheel. J Psychopharmacol. 2007;21(2):134-5.
81. Stanford SC. Open fields (unlike wheels) can be any shape but still miss the target. J Psychopharmacol. 2007;21(2):144.
82. Wahlsten D, Rustay NR, Metten P, Crabbe JC. In search of a better mouse test. Trends Neurosci. 2003;26(3):132-6.
83. Richardson CA. The power of automated behavioural homecage technologies in characterizing disease progression in laboratory mice: A review. Appl Anim Behav Sci. 2015;163(0):19-27.
84. Richter SH. Automated Home-Cage Testing as a Tool to Improve Reproducibility of Behavioral Research? Frontiers in Neuroscience. 2020;14:383.
85. Festing MF. We are not born knowing how to design and analyse scientific experiments. Altern Lab Anim. 2013;41(2):P19-21.
86. Dirnagl U. The p value wars (again). European journal of nuclear medicine and molecular imaging. 2019;46(12):2421-3.